DOCUMENT RESUME

ED 234 087 TM 830 589

AUTHOR Ingebo, George S.; Doherty, Victor W.
TITLE An Empirical Study of the Effects of Regression to the Mean.
INSTITUTION Portland Public Schools, Oreg.
PUB DATE 16 Mar 82
NOTE 25p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Achievement Gains; Elementary Education; Language Arts; Latent Trait Theory; Mathematics Achievement; Reading Achievement; *Regression (Statistics); *Scores; *Test Interpretation
IDENTIFIERS *Portland School District OR

ABSTRACT

The Portland (Oregon) Public Schools undertook studies of regression based on a three-year data base of fall and spring test results in grades 3 through 8 in reading, language usage and mathematics. The objective of these regression studies was to determine the amount of correction for regression to the mean needed to accurately represent the growth from fall to spring of groups of students placed in treatment groups where fall test scores were a principal criterion in the placement of students in those treatment groups. A series of tables are presented to illustrate the findings. Regression patterns for reading, mathematics and language usage were found to be quite similar, although sufficient differences were found to recommend separate tabling of the correction factors for the respective subjects. Based on these results, it appears that statistical theory that shows error of measurement increasing symmetrically toward the low and high ends of a distribution of scores may not represent the reality of measurement error. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

EVALUATION DEPARTMENT
PORTLAND PUBLIC SCHOOLS

PORTLAND, OREGON 97208

ED234087

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

AN EMPIRICAL STUDY OF THE EFFECTS OF REGRESSION TO THE MEAN

George S. Ingebo
Victor W. Doherty

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. S. Ingebo

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Evaluation Department, Portland Public Schools

March 16, 1982

TM 830 589

AN EMPIRICAL STUDY OF THE EFFECTS OF REGRESSION TO THE MEAN

The effects of regression to the mean on achievement measures, while well understood by serious students of statistics and measurement, is given little or no attention by educational evaluators in presenting and interpreting student gain measures for public school programs such as Title I, Title VI ESAA, and other federal program evaluations.

The importance of regression is especially great in these evaluations because the groups receiving treatment normally are selected with test scores as a principal criterion. When students in a treatment group are selected on the basis of low test scores, it normally assures that group of some gain due to regression in its next measurement.

Because the effects of regression have not been empirically established in each of the situations in which different tests and testing practices are used, there is virtually no effort being made to account for regression in studies of student gain.

The magnitude of untruth that is represented in evaluations of Title I programs, for example, is considerable. Decisions are being made every year on the basis of these evaluations which almost universally show more gains than actually occurred because regression has not been taken into account.

The Portland Public Schools, having developed a carefully researched program of measurement and a three-year data base of fall and spring test results in grades 3 through 8, has undertaken a study of regression based on that three-year data base which empirically demonstrates the effects of regression on gains made from fall to spring (as well as fall to fall) in reading, language usage and mathematics.

In order to interpret the results of such a study of regression, it is necessary to understand the character of the testing program that produced the data used in the study. This is especially important in Portland's case because the program of measurement there is different from that in any other city in the United States. It is the only program that has <u>all</u> of the following features:

1. A single Rasch scale extending from low third grade to high eighth grade performance in reading, language usage and mathematics.

2. Carefully established item values tied to this scale by carefully researched and continuously refined linking and calibrating procedures that verify with great accuracy all item value relationships in the test series.

3. Tests graded and overlapping in difficulty to provide every student an opportunity to be measured by a test that efficiently discriminates his/her performance.

4. A placement system that assures proper placement of students in tests that measure them efficiently, and that make it possible for all students in a class to complete their tests in about the same amount of time and with about the same degree of success (items right out of items attempted).

5. Standardized instructions that permit students assigned to different tests to be tested in the same class group.

6. Follow-up retesting of all students who do not perform within the accurate measurement range of the test.

7. Testing of all students in all grades 3 through 8 in mathematics, reading and language usage every fall and spring.

8. Elimination of all scores in the three-year data base where the standard score for a student deviates significantly from the mean of all standard scores for that student over his testing record in a particular subject (to eliminate scores that are not representative for that student).

9. Elimination from the data base of other scores which, for a variety of reasons, must be regarded as inaccurate.

10. Test and test items created to conform precisely to the curriculum of the Portland Public Schools in reading, mathematics and language usage.

11. Elimination of items found to have faulty item characteristics through:

    a. Examination of point biserials;

    b. Examination of mean square fit;

    c. Examination of the performance of large numbers of students in each item in relation to the Rasch item (discrimination) curve;

    d. Elimination of items suspected not to validly represent the domain they were designed to measure. (It should be noted that Rasch methodology permits continuous refinement and improvement of items and tests over time without loss of comparability of the overall measures being employed.)

Effects on Regression of the Specific Character of a Testing Program

Typically, regression to the mean is regarded as a statistical phenomenon independent of the character of a testing program. Logically, however, the character of a testing program should influence the degree of regression in the test results secured by the program. In the case of Portland's functional level testing program, which attempts to place every student within the appropriate measuring range of each test and retests students who do not fall within that range, it seems reasonable to expect that regression would not be as great as in programs where greater measurement error occurs because of the extended range of the tests used, and because students are not retested who score outside the appropriate measurement range of the tests. Also, it would appear logical that tests in which stringent item discrimination criteria are used should have lower measurement error and less regression.

The studies done in Portland could not affirm the logic of these assumptions since the amount of regression could be determined only for Portland's tests. However, empirical studies of regression based on the methodology used in Portland are possible to make in other situations, and such studies would be useful in determining measurement error and regression where other types of tests and testing practices are used.

METHODOLOGY

The objective of these regression studies was to determine as precisely as possible the amount of correction for regression to the mean needed to accurately represent the growth from fall to spring of groups of students placed in treatment groups where fall test scores were a principal criterion in the placement of students in those treatment groups. The following procedure was used:

1. All student scores in grades 3 through 8 were converted to a Portland standard score with a mean of 50 and a standard deviation of 10, based upon the distribution of performance of students at each grade level.

2. Fall standard scores were aggregated across grade levels for each subject to show the number of students scoring at each standard score level.

3. Standard score performance in the spring was pulled from the data base for students scoring at each standard score level in the fall. Mean spring standard scores were then determined for each group. Since student fall scores were aggregated by standard score at each standard score level, the mean for that group in the fall was obviously the standard score level. In the spring, however, the scores of students constituting any standard score level in the fall would produce a distribution with a mean and standard deviation.

4. The standard score for each group aggregated according to that score in the fall was subtracted from the mean standard score performance of that same group of students in the spring, and a distribution was made of the differences between the fall and spring performance for every standard score group. This basically revealed the extent of regression for each standard score group from fall to spring.

5. The method just described was applied first to students in each grade level in each subject to determine if sufficient variation existed from grade level to grade level to require computation of different regression formulas for each grade level, 3 through 8. These studies revealed sufficient similarity in the regression factors that it was decided to combine all grade levels, a procedure made possible by the use of standard scores.

6. Having combined all grade levels, regression factors were computed for each of the three years represented in the data base, and again sufficient similarity was found to suggest the desirability of combining all scores for three years. Thus, by combining across grade levels 3 through 8, and by combining results from three years of testing, large numbers of students could be represented at each standard score level, increasing the reliability of the regression factors arrived at through the study.

7. The distribution of scores in the spring for each fall standard score group yielded a standard deviation that became a measure of the error of predicting spring scores from a fall standard score group. This "error of prediction" was computed to test empirically the assumption of symmetrical variation of error of measurement, as proposed by Benjamin Wright and others.

8. An analysis was made of the skewness within intervals of standard score groupings for which fall to spring comparisons in standard scores were made to test the assumption that this skewness would be decreasingly positive as it moved from low fall P score groups to the mean, and decreasingly negative for high P score groupings as they moved toward the mean. For this analysis, fall standard scores were aggregated by interval groupings of five beginning with a low of 19.5 and progressing to a high of 74.5.

Again, it should be noted that the methodology described above was made possible by the creation of a "clean" data base in which inaccurate and suspect test scores were deleted; by administration of tests in which the discriminating performance of items had been continuously examined and improved over several years; by a program that tests all children in grades 3 through 8 in the fall and spring of every year; and that produces a standard score for each grade level distribution in each subject, both for purposes of reporting results to teachers and to facilitate the type of research represented here.

FINDINGS

Findings of these studies of regression to the mean are summarized in tables A through Q. Tables A and B for reading and mathematics, respectively, show the effects of regression in standard score terms for students whose fall standard scores ranged from 30 to 70. The broken line in these tables represents the actual regression line based on the amount of regression found at each standard score point along the scale. The straight line represents the points at which correction in the regression factor is required for each half of each standard score unit of regression change.

This information is provided in tabular form in Tables C and D for reading and mathematics, respectively. These are the tables from which corrections to gain scores from fall to spring will be made in evaluating the effects of regression to the mean on fall to spring achievement gains of groups selected for treatment on the basis of fall test scores. The tables are quite similar for the two subjects, except that a somewhat greater range of midpoint scores require no correction in reading than in math, with a little greater correction being required in mathematics for low scores, and a little greater correction in reading for high scores. It should be noted that the correction of group means in the magnitudes shown could have sufficient influence on gains reported for treatment groups to change interpretation of the significance of these gains.

Table E shows the effects of regression to the mean for reading and mathematics and reveals a close similarity in the regression factors for these two subjects. Note that for these initial tables, large numbers of students are represented even at the extremes of the standard score scale due to the combining of standard scores across grade levels and for three different years.

Some indication of the magnitude of the numbers can be secured by examining Table Q, which shows these numbers for only one of the three fall to spring score groups. Tripling these numbers would approximate the numbers reflected in graphs A through E for each subject.

Tables F through K show the regression factors for math and reading in each of the years 1978-79, 1979-80 and 1980-81. These tables are presented to show the basic similarity of the regression data across the years which led to the decision to combine data for all three years in arriving at the correction factors. (In examining these tables, it should be noted that some span a standard score range of 25 to 75, whereas others cover a span of 30 to 70. This occurred simply because the graphs were made at different times.

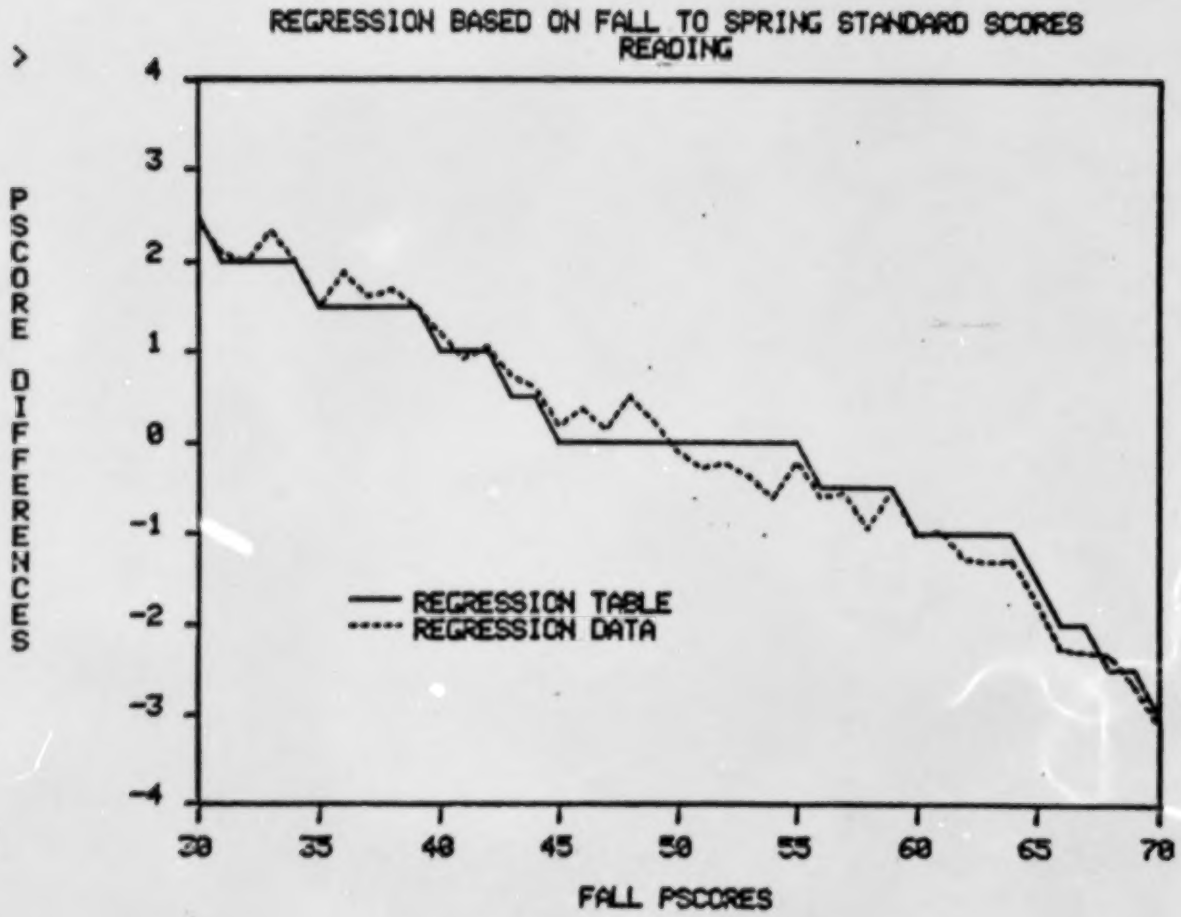
Tables L through O graphically illustrate the regression phenomenon with fall P score group means plotted along the X axis and spring standard scores plotted along the Y axis. Table N introduces data from a similar study made of language usage test scores, and Table O plots the regression of all three subjects on the same graph. This table shows somewhat greater regression at the extremes for language usage than was found for reading and mathematics.

Table P represents data relating to the area of prediction of spring scores from fall scores. Where the preceding regression tables showed the movement of P score means from fall to spring, this table shows the extent to which each standard score group in the fall expanded as a distribution of scores in the spring. For example, if 500 students had a standard score (P score) of 25 in the fall, the distribution of scores in the spring for those same students might have a standard deviation of 6, whereas a standard score group of 70 in the fall may have a distribution of scores in the spring with a standard deviation of only 4. This would show that prediction of the spring scores from the fall score is less accurate for the P score group of 30 than it is for the P score group of 70. This procedure possibly produces a better measure of error of measurement than the standard statistical procedure for producing a Rasch error of measurement estimation. Formulas used for this purpose tend to show increasing error as scores move from the mean to both extremes; data in Table P indicate that error of measurement declines from low scores to high scores, with only a slight upturn in error in the 65 to 75 standard score range.

Table Q, to which reference has been made earlier, simply confirms through use of a statistical measure of skewness that the standard scores represented have a systematic progression of skewness from interval to interval and in the directions that would be expected. The formula used to determine an index of skewness is as follows:

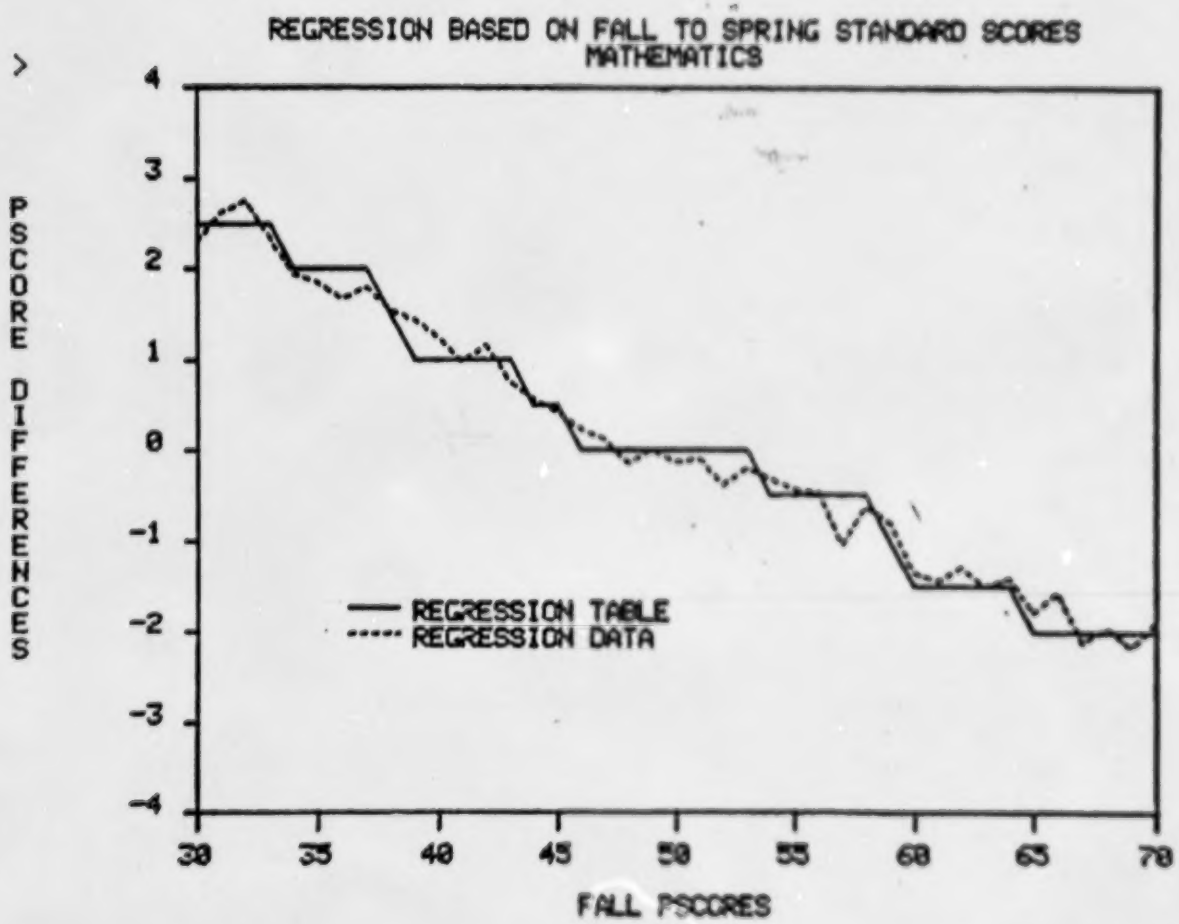
$$\text{Skewness} = \frac{\sum_{i=1}^{N} [(X_i - \bar{X})/s]^3}{N}$$

TABLE A



Combines data for 78-79, 79-80, 80-81

TABLE B



Combines data for 78-79, 79-80, 80-81

TABLE C

TABLE FOR CORRECTING
SPRING GROUP MEANS
(To be used in interpreting growth from fall
to spring of groups selected for treatment
on the basis of fall test scores.)

```
LIST

30 2.5
31 2
32 2
33 2
34 2
35 1.5
36 1.5
37 1.5
38 1.5
39 1.5
40 1
41 1
42 1
43 .5
44 .5
45 0
46 0
47 0
48 0
49 0
50 0
51 0
52 0
53 0
54 0
55 0
56 -.5
57 -.5
58 -.5
59 -.5
60 -1
61 -1
62 -1
63 -1
64 -1
65 -1.5
66 -2
67 -2
68 -2.5
69 -2.5
70 -3

#READING REGRESSION TABLE
```

10

TABLE D

TABLE FOR CORRECTING
SPRING GROUP MEANS
(To be used in interpreting growth from fall
to spring of groups selected for treatment
on the basis of fall test scores.)

LIST

30	2.5
31	2.5
32	2.5
33	2.5
34	2
35	2
36	2
37	2
38	1.5
39	1.
40	1
41	1
42	1
43	1
44	.5
45	.5
46	0
47	0
48	0
49	0
50	0
51	0
52	0
53	0
54	-.5
55	-.5
56	-.5
57	-.5
58	-.5
59	-1
60	-1.5
61	-1.5
62	-1.5
63	-1.5
64	-1.5
65	-2
66	-2
67	-2
68	-2
69	-2
70	-2

#MATHEMATICS REGRESSION TABLE

11

TABLE E

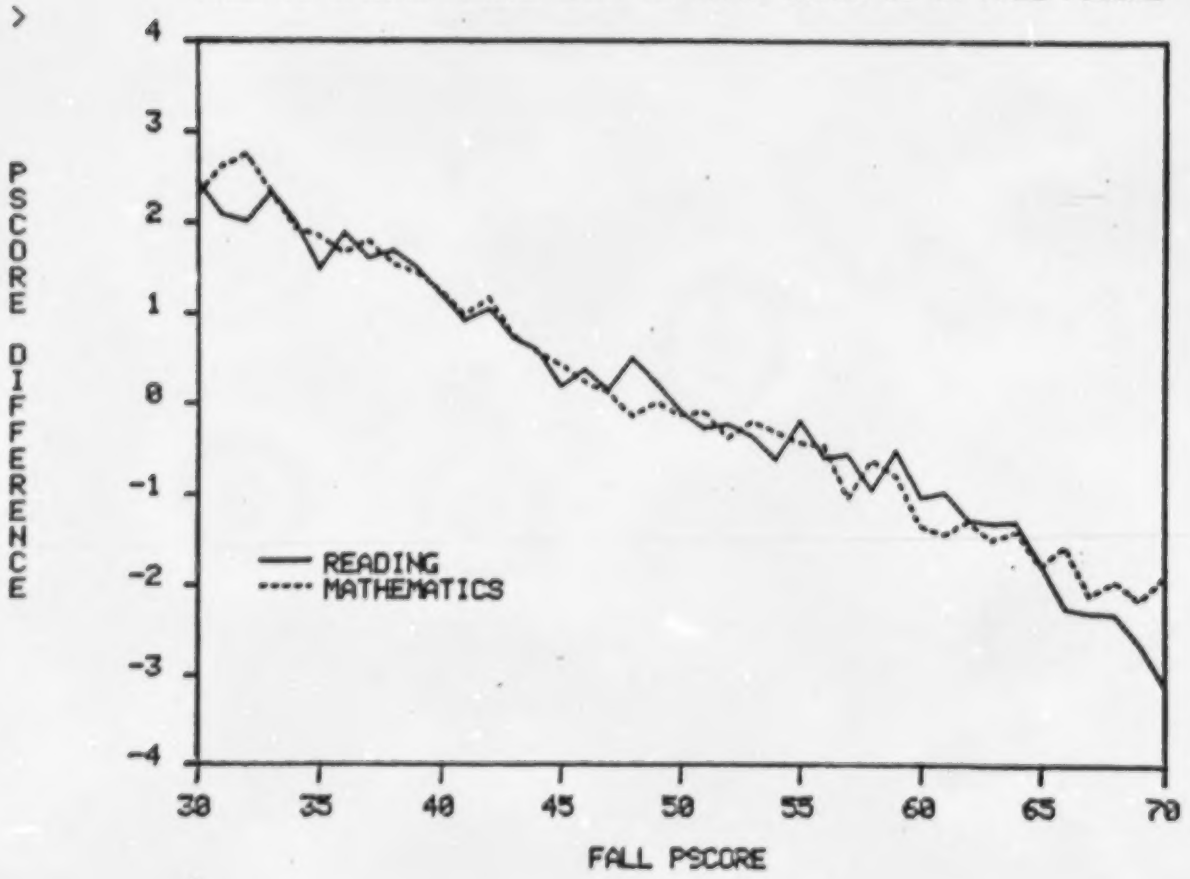

FALL TO SPRING PSCORE GAIN 78-79,79-80,80-81 BY FALL PSCORE
PSCORE DIFFERENCE
4
3
2
1
0
-1
-2
-3
-4
READING
MATHEMATICS
30
35
40
45
50
55
60
65
70
FALL PSCORE

TABLE F

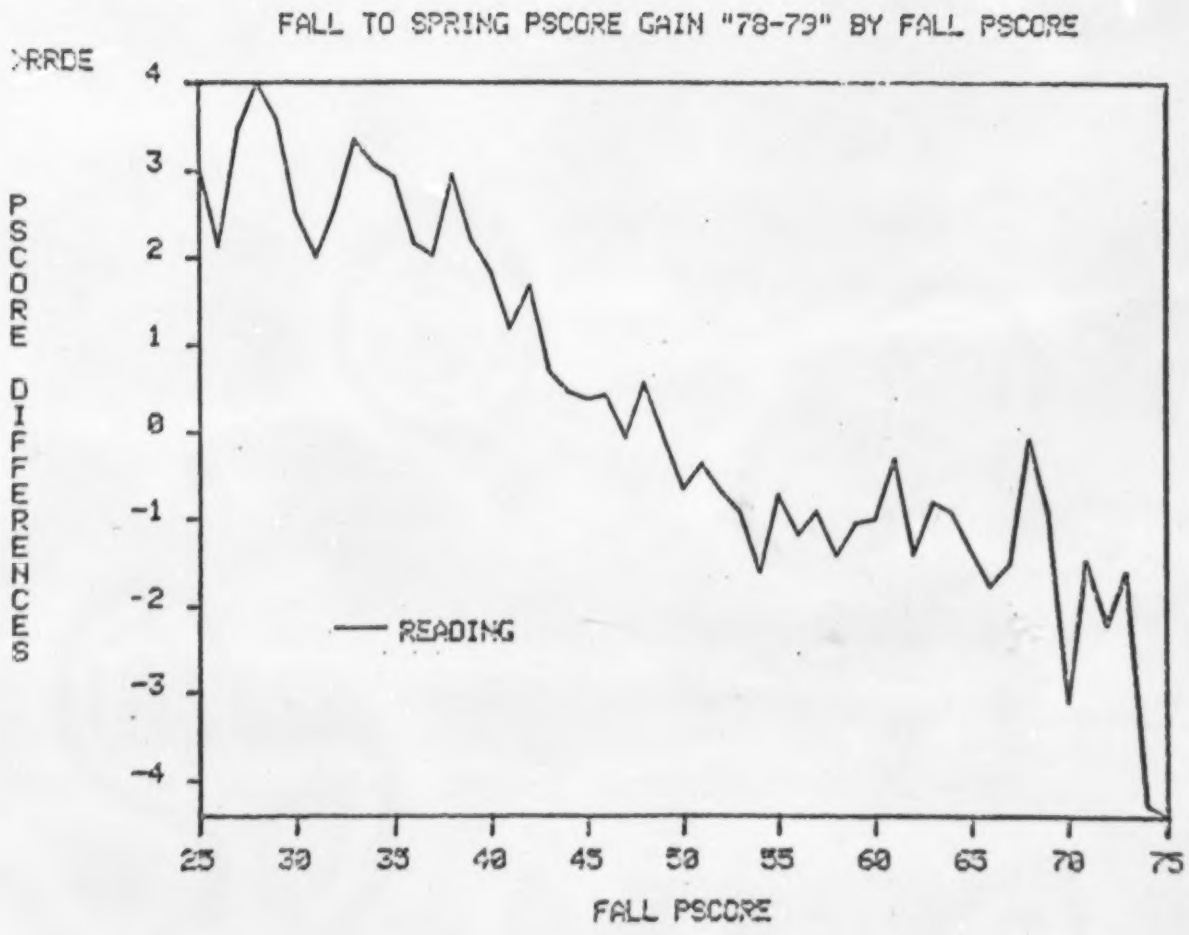

FALL TO SPRING PSCORE GAIN "78-79" BY FALL PSCORE
PSCORE DIFFERENCES
4
3
2
1
0
-1
-2
-3
-4
READING
25
30
35
40
45
50
55
60
65
70
75
FALL PSCORE

TABLE G

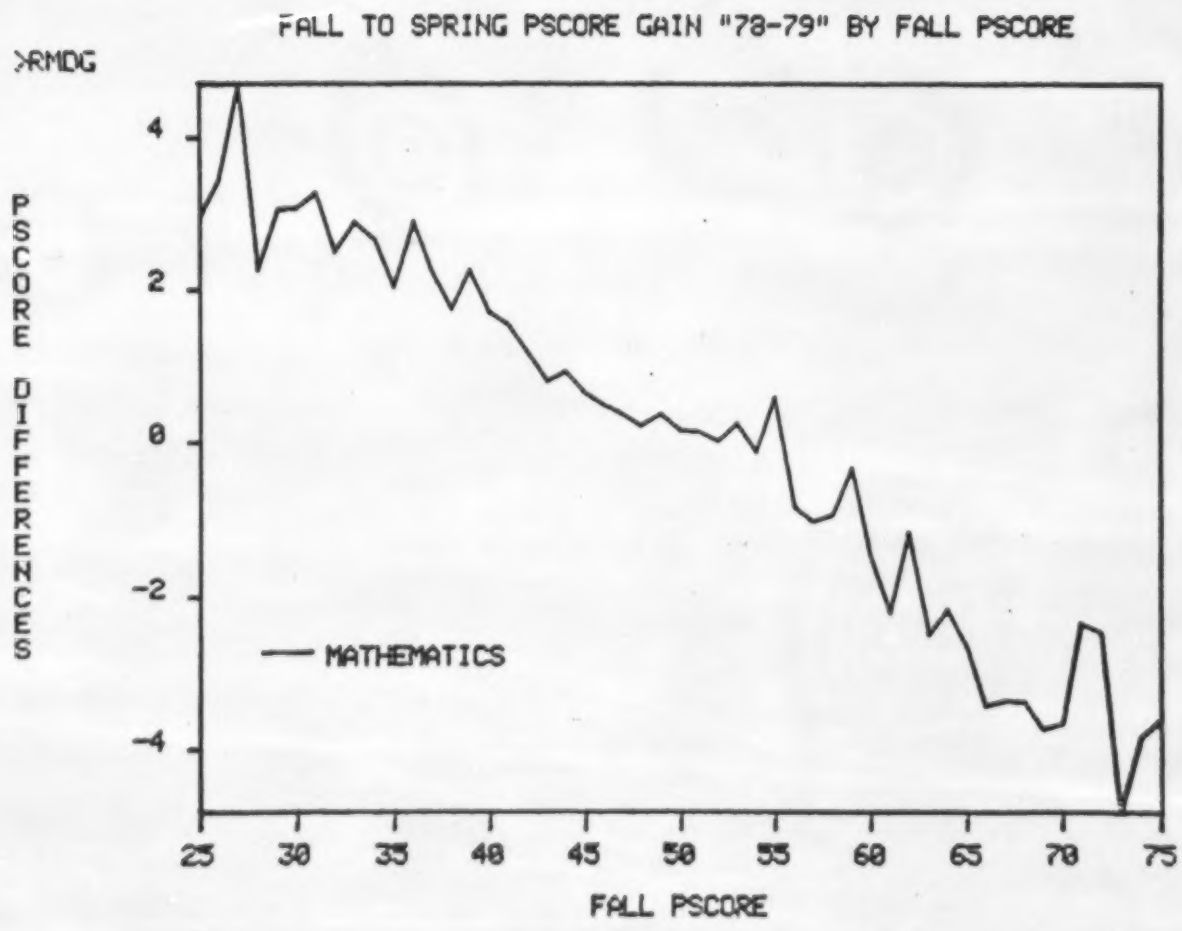

FALL TO SPRING PSCORE GAIN "78-79" BY FALL PSCORE
>RMDG
PSCORE DIFFERENCES
4
2
0
-2
-4
MATHEMATICS
25
30
35
40
45
50
55
60
65
70
75
FALL PSCORE

TABLE H

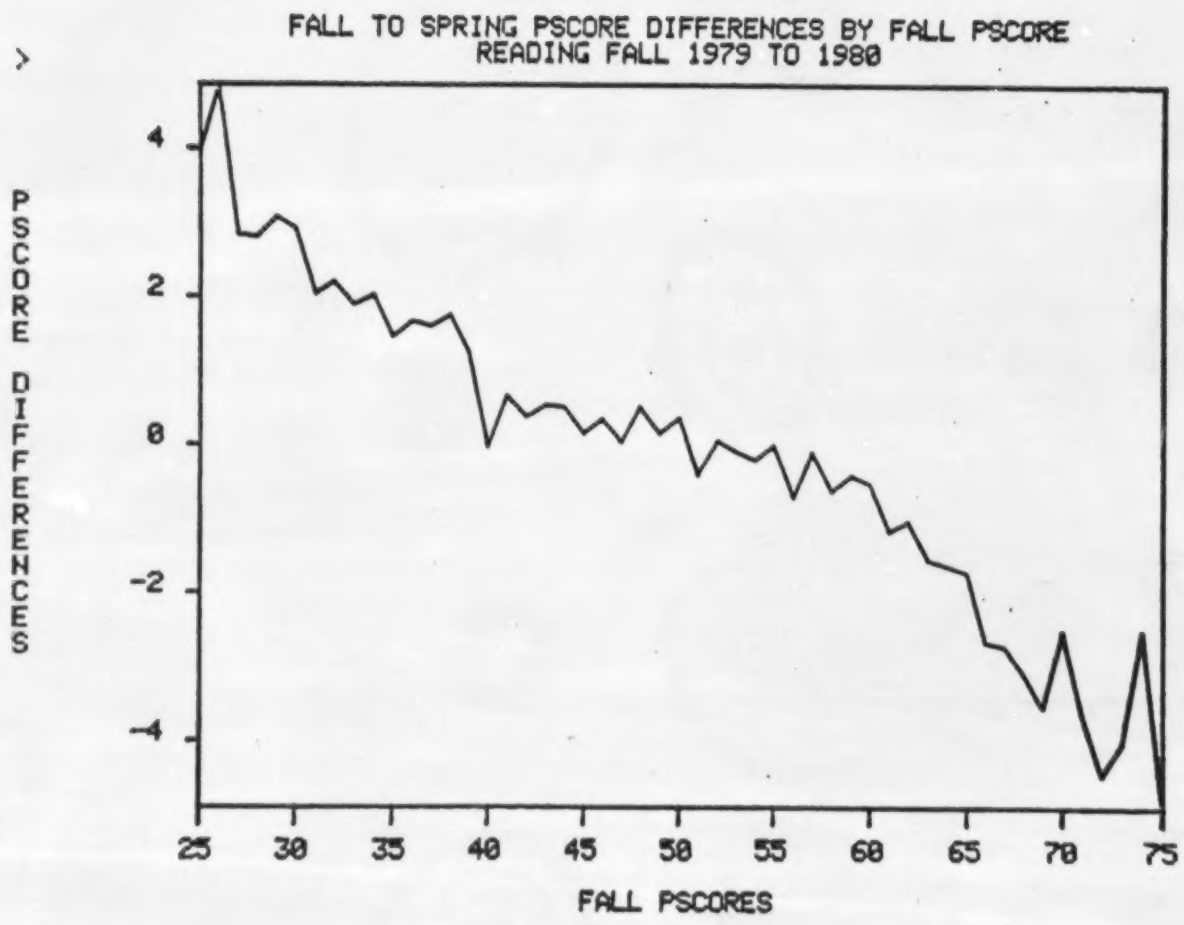

FALL TO SPRING PSCORE DIFFERENCES BY FALL PSCORE
READING FALL 1979 TO 1980
PSCORE DIFFERENCES
4
2
0
-2
-4
25
30
35
40
45
50
55
60
65
70
75
FALL PSCORES

TABLE I

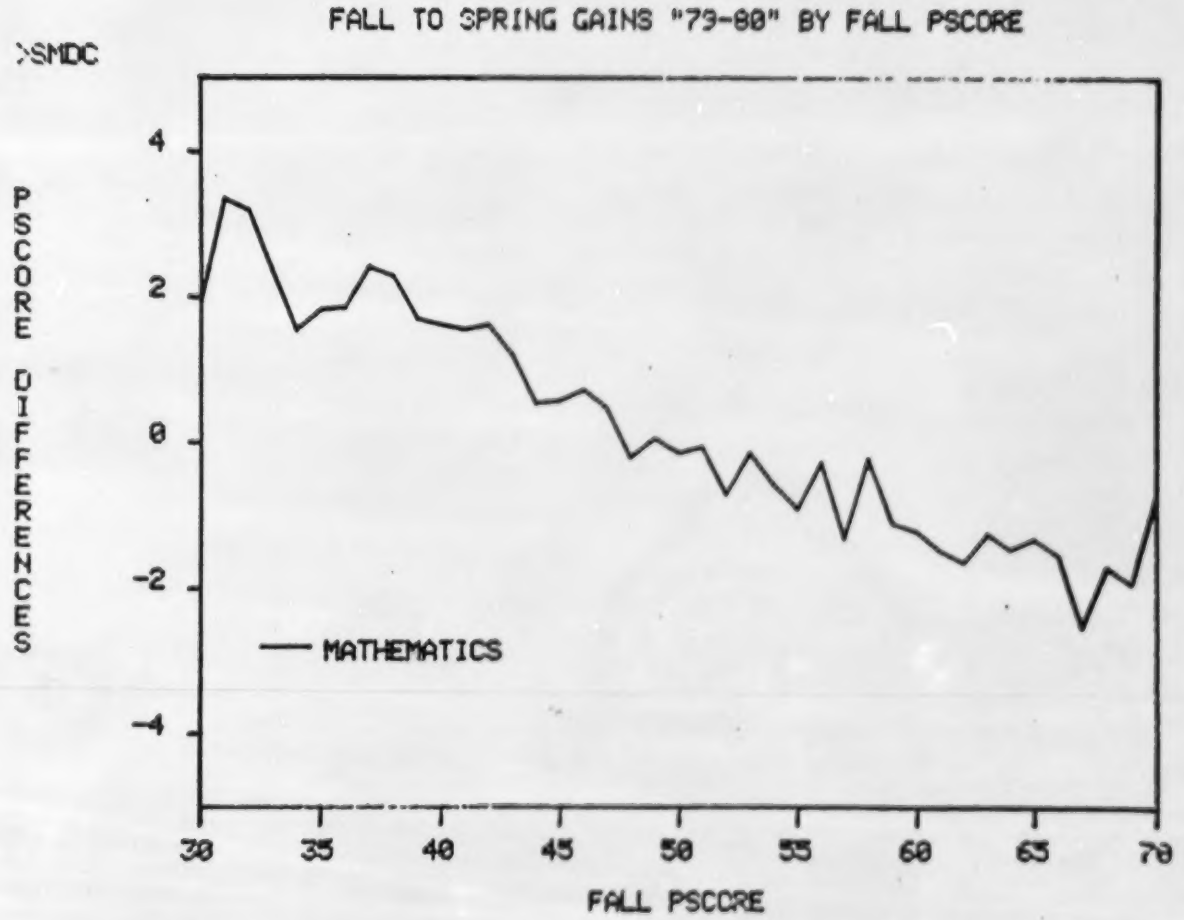

FALL TO SPRING GAINS "79-80" BY FALL PSCORE
>SMDC
PSCORE DIFFERENCES
4
2
0
-2
-4
MATHEMATICS
30
35
40
45
50
55
60
65
70
FALL PSCORE

TABLE J

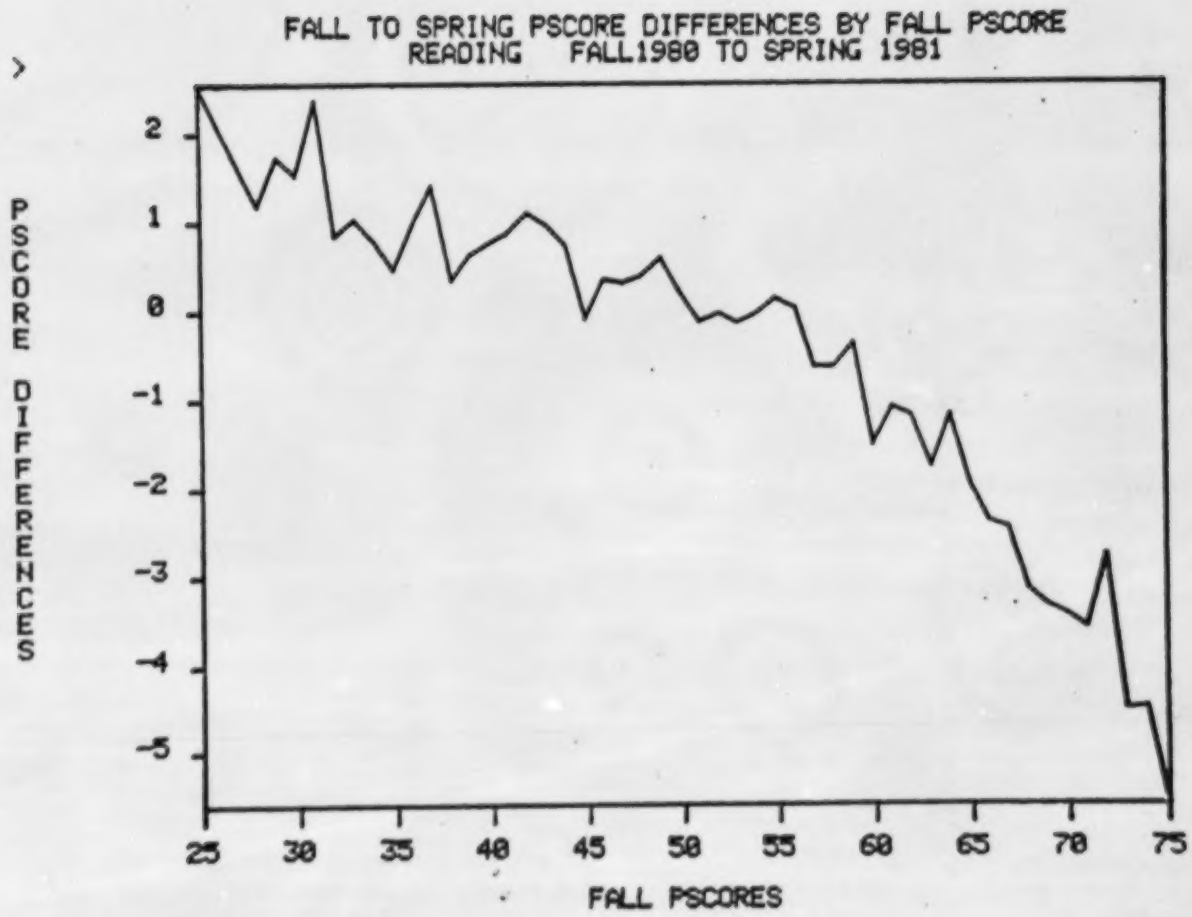

FALL TO SPRING PSCORE DIFFERENCES BY FALL PSCORE
READING FALL1980 TO SPRING 1981
PSCORE DIFFERENCES
2
1
0
-1
-2
-3
-4
-5
25
30
35
40
45
50
55
60
65
70
75
FALL PSCORES

TABLE K

FALL TO SPRING GAINS "80-81" BY FALL 1980 PSCORE

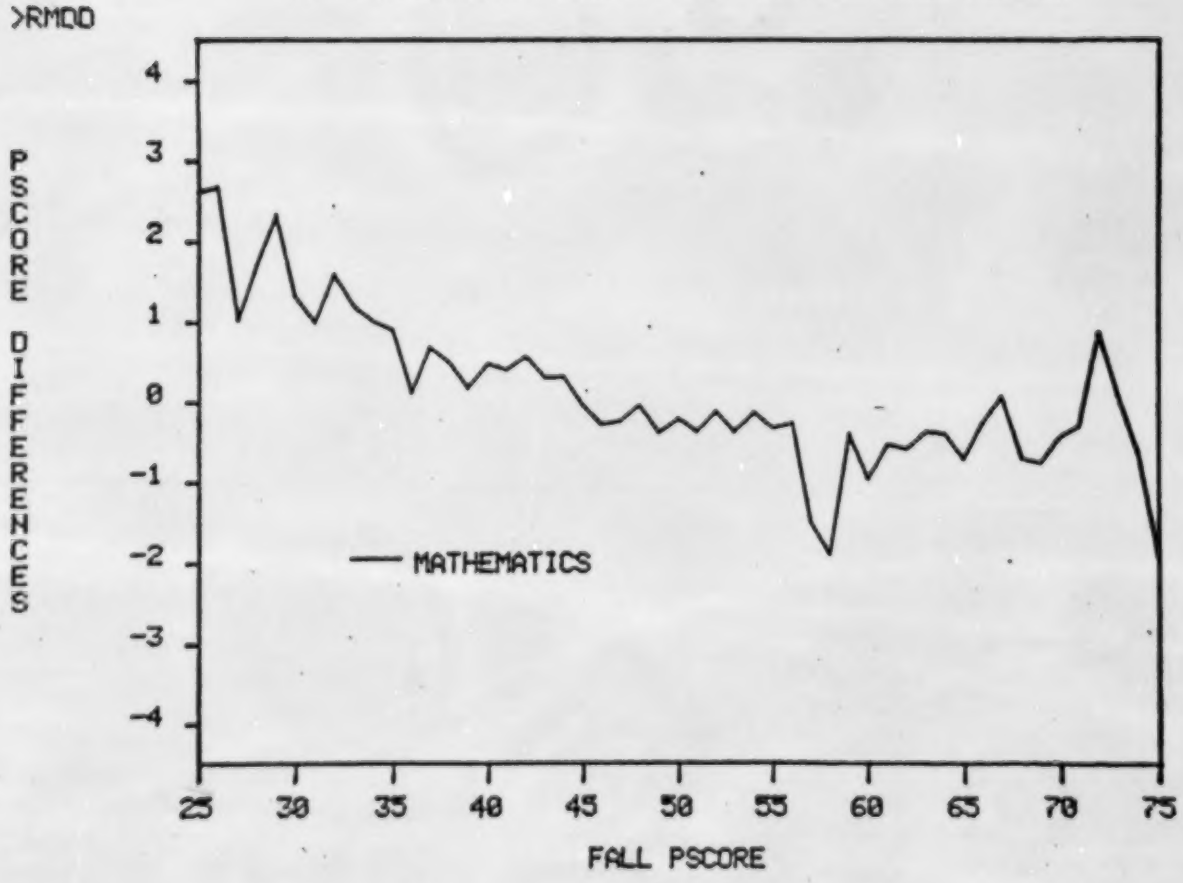
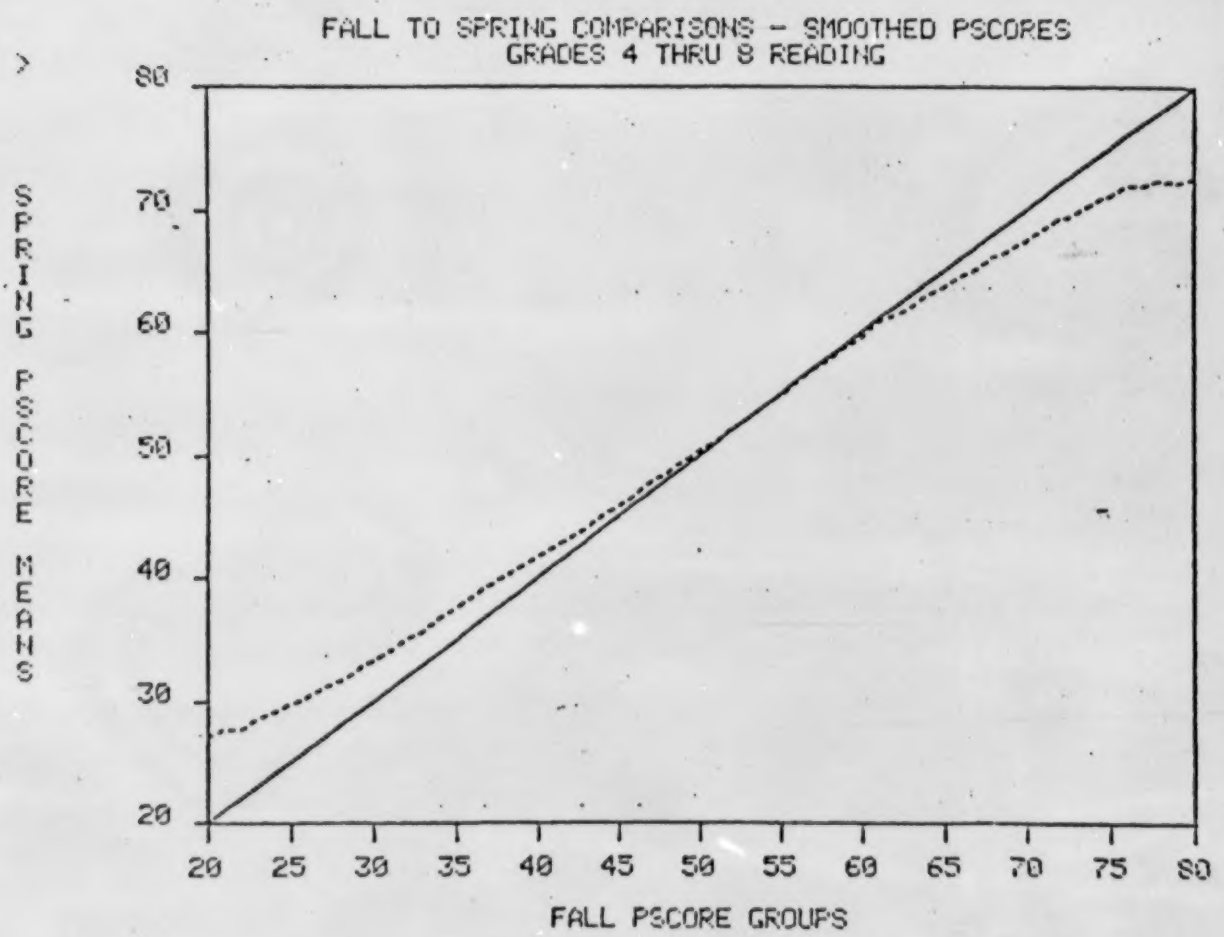
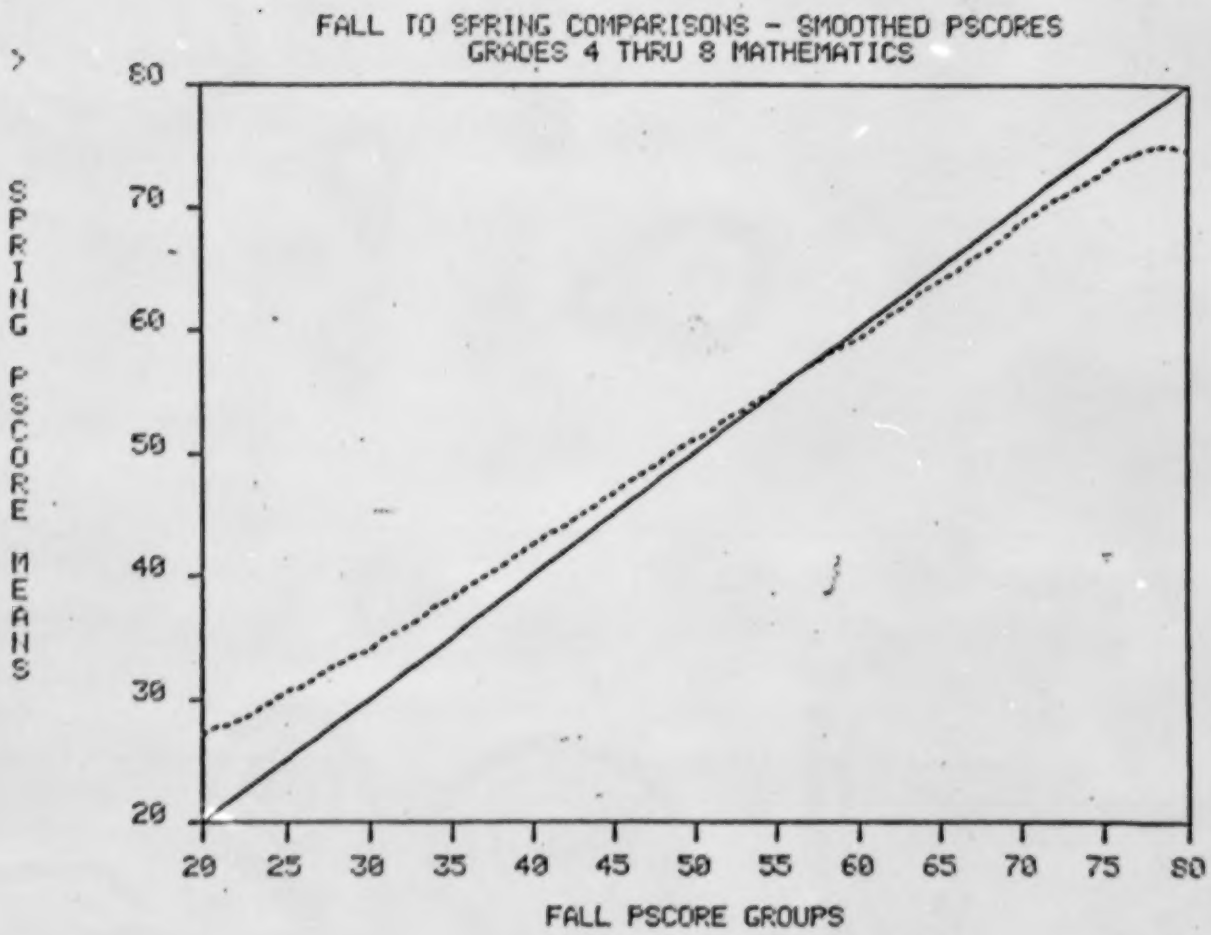


TABLE L



Combines data for 78-79, 79-80, 80-81

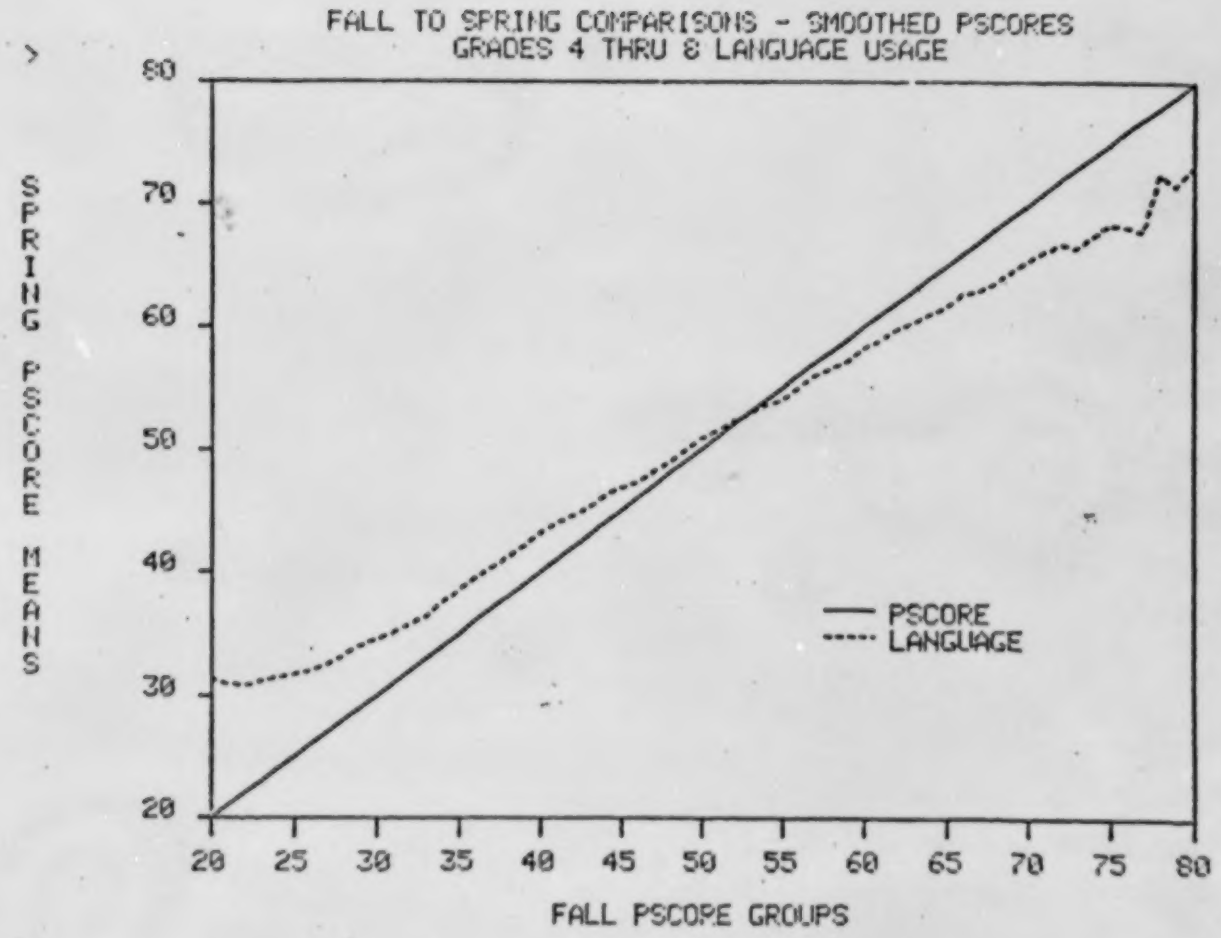
TABLE M



Combines data for 78-79, 79-80, 80-81

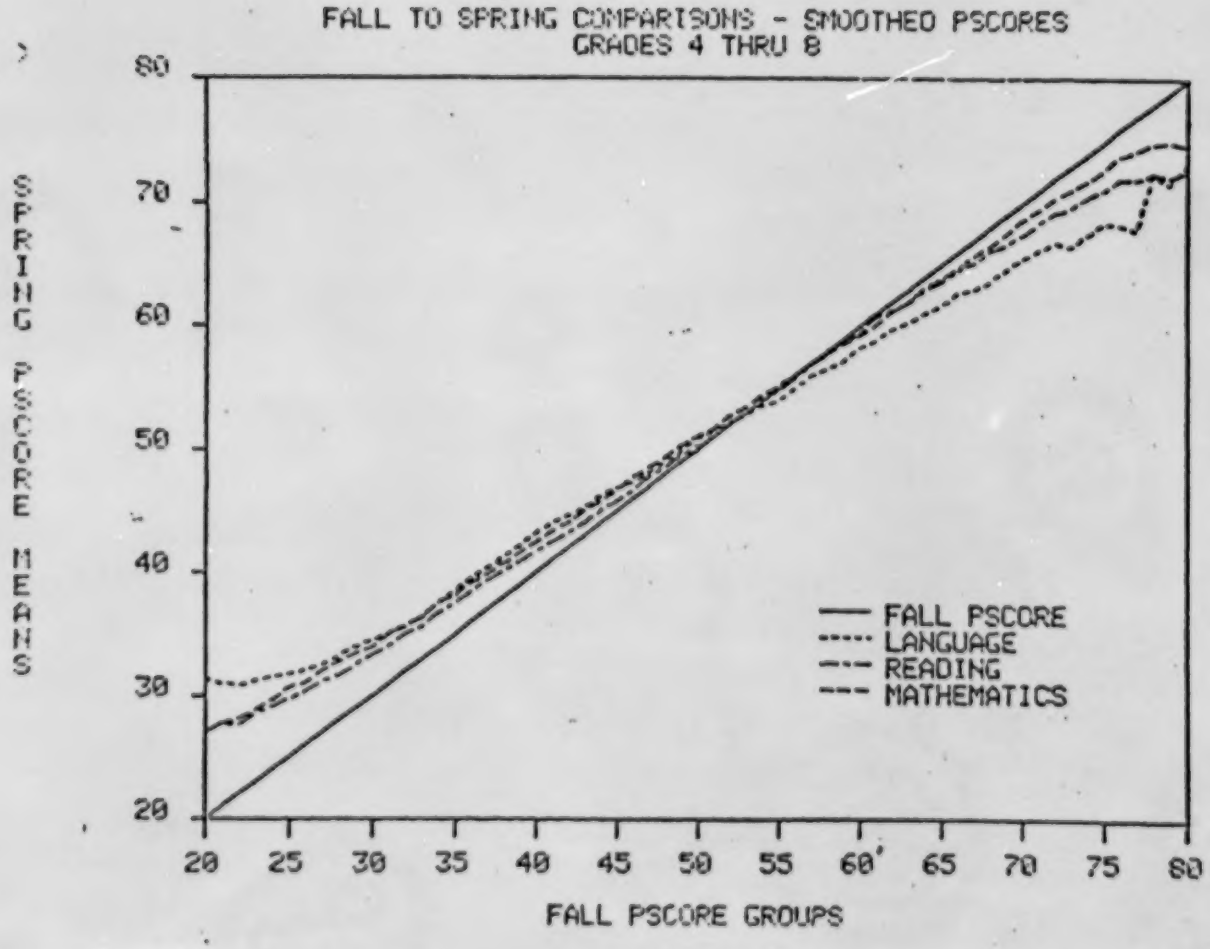
4/21/81

TABLE N



Combines data for 78-79, 79-80, 80-81

TABLE 0

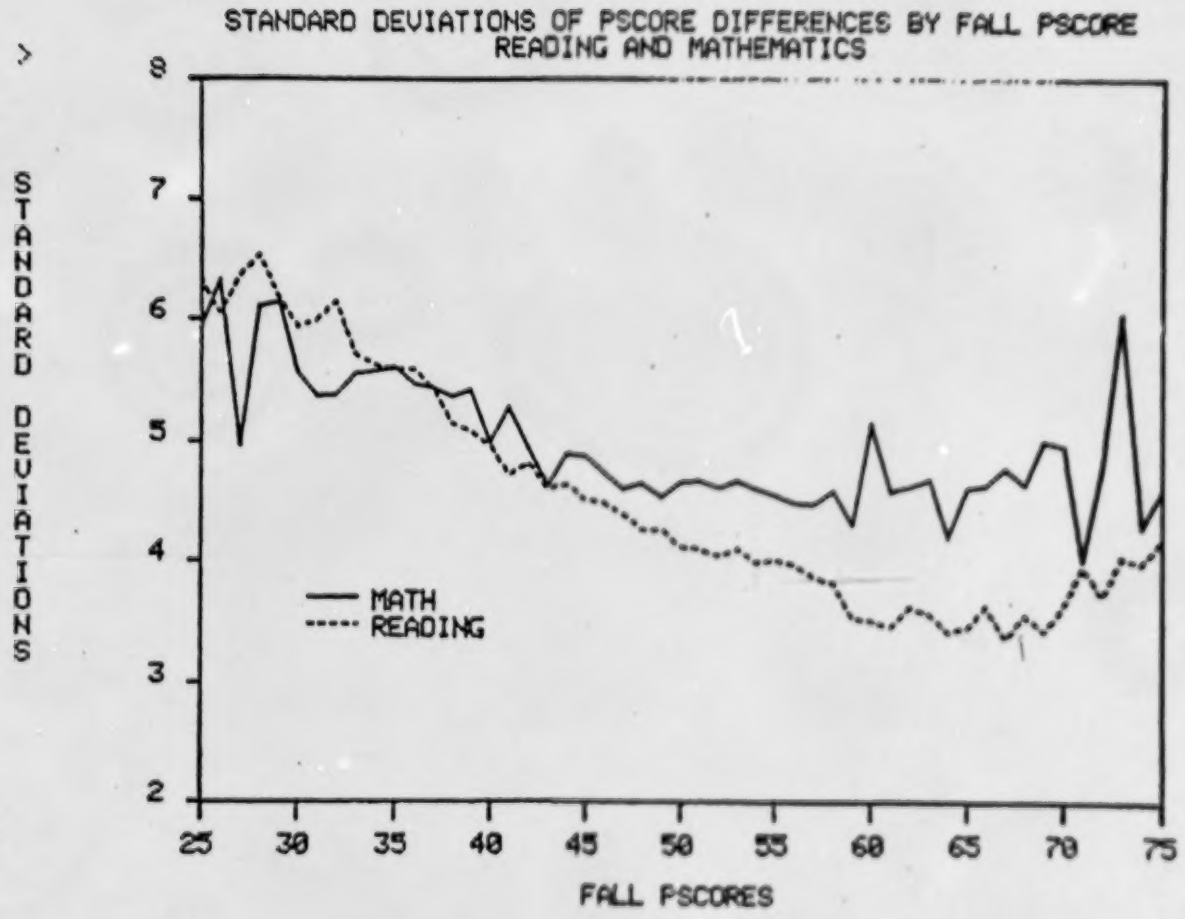


Combines data for 78-79, 79-80, 80-81

22

4/21/81

TABLE P



Combines data for 78-79, 79-80, 80-81

TABLE Q

Means and Skewness of Distributions of Fall (1979) to Spring (1980)
Standard Score Differences Grouped by Fall (1979)
Standard Score in Intervals of Half Standard Deviation

READING

P-Scores	Fall P-Scores		Spring Minus Fall Differences				Number
	X	SK	X	SK	Min.	Max.	
20	23.4	-.48	5.01	.73	- 5	23	162
25	28.0	-.35	1.94	.63	-10	25	497
30	32.8	-.34	.47	.38	-15	25	983
35	37.8	-.18	-.21	- .01	-19	22	1640
40	42.7	-.21	-.30	- .16	-22	18	2920
45	47.7	-.09	-.42	- .36	-25	21	4387
50	52.5	.01	-.38	- .11	-21	21	4890
55	57.2	.21	-.40	- .06	-23	17	3983
60	62.1	.43	-1.08	.43	-20	14	2200
65	66.9	.48	-2.08	+ .34	-20	13	890
70	71.6	.63	-3.30	+ .42	-16	9	179

MATHEMATICS

20	23.0	-.35	8.00	.60	- 4	25	153
25	28.2	-.53	3.42	.55	-10	23	417
30	32.8	-.34	1.58	.31	-16	21	875
35	37.9	-.23	.24	.18	-19	21	1756
40	42.7	-.23	-.04	.15	-22	21	3163
45	47.7	-.08	-.74	- .03	-22	20	4493
50	52.5	.02	-.86	- .05	-21	18	4775
55	57.2	.18	-1.14	- .15	-24	17	3827
60	61.9	.41	-1.39	- .12	-24	17	1960
65	67.0	.51	-1.58	- .39	-23	10	762
70	71.7	.63	-2.43	- .55	-16	6	199

LANGUAGE USAGE

20	23.4	-.69	4.01	.77	- 5	24	157
25	28.0	-.48	2.03	.53	- 9	23	390
30	33.0	-.37	1.12	.37	-14	23	1026
35	37.9	-.32	.42	.08	-18	25	1745
40	42.7	-.18	.23	.11	-26	25	2891
45	47.8	-.15	-.15	- .30	-26	17	4283
50	52.5	.06	-.43	- .30	-23	15	5072
55	57.3	.12	-1.20	- .30	-22	12	4098
60	61.9	.40	-2.15	- .17	-19	10	2237
65	67.1	.36	-3.68	- .19	-23	11	705
70	71.3	1.00	-5.32	- .53	-22	5	216

3/9/82
GI/wp

CONCLUSIONS

The methods employed for determining the extent of regression to the mean produced numerical estimates that clearly demonstrate the amount of regression encountered in Portland's measurement program. Using the information produced, it will be possible to make corrections in gain measures for groups selected for treatment on the basis of fall test scores. Regression patterns for reading, mathematics and language usage were found to be quite similar, although sufficient differences were found to recommend separate tabling of the correction factors for the respective subjects. The data on error of prediction from fall to spring scores revealed similar patterns for math and reading. In both subjects, the error of prediction decreased from low to high scores. Assuming that the procedure used yields a reasonable approximation to error of measurement, it appears that statistical theory that shows error of measurement increasing symmetrically toward the low and high ends of a distribution of scores may not in fact represent the reality of measurement error, at least for the type of testing program developed in Portland.

VWD/wp

3/16/82